SPIDER-DAY03

1. lxml解析库

1.1 安装使用流程

```
【1】安装
   sudo pip3 install lxml

【2】使用流程
   2.1》导模块 :           from lxml import etree
   2.2》创建解析对象 :      parse_html = etree.HTML(html)
   2.3》解析对象调用xpath : r_list = parse_html.xpath('xpath表达式')

【此生铭记】: 只要调用了xpath，得到的结果一定是列表
```

1.2 lxml+xpath使用

```
【1】基准xpath: 匹配所有电影信息的节点对象列表
  //dl[@class="board-wrapper"]/dd
  [<element dd at xxx>,<element dd at xxx>,...]

【2】遍历对象列表，依次获取每个电影信息
  item = {}
  for dd in dd_list:
       item['name'] = dd.xpath('.//p[@class="name"]/a/text()').strip()
       item['star'] = dd.xpath('.//p[@class="star"]/text()').strip()
       item['time'] = dd.xpath('.//p[@class="releasetime"]/text()').strip()
```

2. 豆瓣图书爬虫

2.1 需求分析

```
【1】抓取目标 - 豆瓣图书top250的图书信息
    https://book.douban.com/top250?start=0
    https://book.douban.com/top250?start=25
    https://book.douban.com/top250?start=50
    ... ...

【2】抓取数据
    2.1) 书籍名称 : 红楼梦
    2.2) 书籍描述 : [清] 曹雪芹 著 / 人民文学出版社 / 1996-12 / 59.70元
    2.3) 书籍评分 : 9.6
    2.4) 评价人数 : 286382人评价
    2.5) 书籍类型 : 都云作者痴，谁解其中味?
```

2.2 实现流程

```
【1】确认数据来源 - 响应内容存在
【2】分析URL地址规律 - start为0 25 50 75 ...
【3】xpath表达式
    3.1) 基准xpath,匹配每本书籍的节点对象列表
         //div[@class="indent"]/table

    3.2) 依次遍历每本书籍的节点对象，提取具体书籍数据
        书籍名称 :  .//div[@class="pl2"]/a/@title
        书籍描述 :  .//p[@class="pl"]/text()
        书籍评分 :  .//span[@class="rating_nums"]/text()
        评价人数 :  .//span[@class="pl"]/text()
        书籍类型 :  .//span[@class="inq"]/text()
```

2.3 代码实现

```python
import requests
from lxml import etree
import time
import random
from fake_useragent import UserAgent

class DoubanBookSpider:
    def __init__(self):
        self.url = 'https://book.douban.com/top250?start={}'

    def get_html(self, url):
        headers = { 'User-Agent':UserAgent().random }
        html = requests.get(url=url, headers=headers).content.decode('utf-8','ignore')
        # 直接调用解析函数
        self.parse_html(html)

    def parse_html(self, html):
        p = etree.HTML(html)
```

```
        # 基准xpath,匹配每本书的节点对象列表
        table_list = p.xpath('//div[@class="indent"]/table')
        for table in table_list:
            item = {}
            # 书名
            name_list = table.xpath('.//div[@class="pl2"]/a/@title')
            item['book_name'] = name_list[0].strip() if name_list else None
            # 信息
            info_list = table.xpath('.//p[@class="pl"]/text()')
            item['book_info'] = info_list[0].strip() if info_list else None
            # 评分
            score_list = table.xpath('.//span[@class="rating_nums"]/text()')
            item['book_score'] = score_list[0].strip() if score_list else None
            # 人数
            number_list = table.xpath('.//span[@class="pl"]/text()')
            item['book_number'] = number_list[0].strip()[1:-1].strip() if number_list else None
            # 描述
            comment_list = table.xpath('.//span[@class="inq"]/text()')
            item['book_comment'] = comment_list[0].strip() if comment_list else None

            print(item)

    def run(self):
        for i in range(10):
            start = i * 25
            page_url = self.url.format(start)
            self.get_html(url=page_url)
            # 控制数据抓取的频率,uniform生成指定范围内浮点数
            time.sleep(random.uniform(0, 3))


if __name__ == '__main__':
    spider = DoubanBookSpider()
    spider.run()
```

3. 百度贴吧小视频爬虫

3.1 需求分析

```
【1】官网地址：进入某个百度贴吧，寻找有视频的帖子，比如如下帖子链接：
   https://tieba.baidu.com/p/7185877941

【2】目标
   2.1> 在此帖子中提取中具体视频的链接(src)
   2.2> 将视频抓取保存到本地文件(向src发请求获取bytes数据类型,以wb方式保存到本地)
```

3.2 实现流程

```
【1】确认数据来源 ： 静态！！！
【2】帖子中视频的xpath表达式

### 重要：页面中xpath不能全信，一切以响应内容为主
### 重要：页面中xpath不能全信，一切以响应内容为主
### 重要：页面中xpath不能全信，一切以响应内容为主
### 重要：页面中xpath不能全信，一切以响应内容为主
### 重要：页面中xpath不能全信，一切以响应内容为主
### 重要：页面中xpath不能全信，一切以响应内容为主
### 重要：页面中xpath不能全信，一切以响应内容为主
### 重要：页面中xpath不能全信，一切以响应内容为主
### 重要：页面中xpath不能全信，一切以响应内容为主
### 重要：页面中xpath不能全信，一切以响应内容为主
```

3.3 代码实现

```python
import requests
from lxml import etree

# 向帖子链接发请求
url = 'https://tieba.baidu.com/p/7185877941'
headers = {'User-Agent':'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/87.0.4280.88 Safari/537.36'}
html = requests.get(url=url, headers=headers).text

# 视频链接的xpath表达式 - 一切以响应内容为准
x = '//div[@class="video_src_wrapper"]/embed/@data-video'
eobj = etree.HTML(html)
video_url_list = eobj.xpath(x)

# 将视频抓取保存到本地文件
if video_url_list:
    video_url = video_url_list[0]
    video_html = requests.get(url=video_url, headers=headers).content
    with open('girl.mp4', 'wb') as f:
        f.write(video_html)
else:
    print('提取视频链接失败')
```

4. 代理参数

4.1 代理IP概述

```
【1】定义
    代替你原来的IP地址去对接网络的IP地址

【2】作用
    隐藏自身真实IP,避免被封

【3】获取代理IP网站
    快代理、全网代理、代理精灵、... ...

【4】参数类型
    proxies
    proxies = { '协议':'协议://IP:端口号' }
    proxies = { '协议':'协议://用户名:密码@IP:端口号' }
```

4.2 代理分类

4.2.1 普通代理

```
【1】代理格式
    proxies = { '协议':'协议://IP:端口号' }

【2】使用免费普通代理IP访问测试网站: http://httpbin.org/get

import requests
url = 'http://httpbin.org/get'
headers = {'User-Agent':'Mozilla/5.0'}
# 定义代理,在代理IP网站中查找免费代理IP
proxies = {
    'http':'http://112.85.164.220:9999',
    'https':'https://112.85.164.220:9999'
}
html = requests.get(url,proxies=proxies,headers=headers,timeout=5).text
print(html)
```

4.2.2 私密代理和独享代理

```
【1】代理格式
    proxies = { '协议':'协议://用户名:密码@IP:端口号' }

【2】使用私密代理或独享代理IP访问测试网站: http://httpbin.org/get

import requests
url = 'http://httpbin.org/get'
proxies = {
    'http': 'http://309435365:szayclhp@106.75.71.140:16816',
    'https':'https://309435365:szayclhp@106.75.71.140:16816',
}
headers = {
    'User-Agent' : 'Mozilla/5.0',
}
```

```
15
16 html = requests.get(url,proxies=proxies,headers=headers,timeout=5).text
17 print(html)
```

4.3 建立代理IP池

```
1  """
2  建立开放代理的代理ip池
3  """
4  import requests
5
6  class ProxyPool:
7      def __init__(self):
8          self.api_url = 'http://dev.kdlapi.com/api/getproxy/?
   orderid=999955248138592&num=20&protocol=2&method=2&an_ha=1&sep=1'
9          self.test_url = 'http://httpbin.org/get'
10         self.headers = {'User-Agent':'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
   AppleWebKit/537.36 (KHTML, like Gecko) Chrome/85.0.4183.83 Safari/537.36'}
11
12     def get_proxy(self):
13         html = requests.get(url=self.api_url, headers=self.headers).text
14         # proxy_list: ['1.1.1.1:8888','2.2.2.2:9999,...]
15         proxy_list = html.split('\r\n')
16         for proxy in proxy_list:
17             # 测试proxy是否可用
18             self.test_proxy(proxy)
19
20     def test_proxy(self, proxy):
21         """测试1个代理ip是否可用"""
22         proxies = {
23             'http' : 'http://{}'.format(proxy),
24             'https': 'https://{}'.format(proxy),
25         }
26         try:
27             resp = requests.get(url=self.test_url, proxies=proxies, headers=self.headers,
   timeout=3)
28             if resp.status_code == 200:
29                 print(proxy,'\033[31m可用\033[0m')
30             else:
31                 print(proxy,'不可用')
32         except Exception as e:
33             print(proxy, '不可用')
34
35     def run(self):
36         self.get_proxy()
37
38 if __name__ == '__main__':
39     spider = ProxyPool()
40     spider.run()
```

5. requests.post()

5.1 POST请求

```
【1】适用场景 ： Post类型请求的网站

【2】参数 ： data={}
  2.1) Form表单数据：字典
  2.2) res = requests.post(url=url,data=data,headers=headers)

【3】POST请求特点 ： Form表单提交数据
```

5.2 控制台抓包

- **打开方式及常用选项**

```
【1】打开浏览器，F12打开控制台，找到Network选项卡

【2】控制台常用选项
  2.1) Network：抓取网络数据包
    a> ALL：抓取所有的网络数据包
    b> XHR：抓取异步加载的网络数据包
    c> JS ： 抓取所有的JS文件
  2.2) Sources：格式化输出并打断点调试JavaScript代码，助于分析爬虫中一些参数
  2.3) Console：交互模式，可对JavaScript中的代码进行测试

【3】抓取具体网络数据包后
  3.1) 单击左侧网络数据包地址，进入数据包详情，查看右侧
  3.2) 右侧：
    a> Headers：整个请求信息
       General、Response Headers、Request Headers、Query String、Form Data
    b> Preview：对响应内容进行预览
    c> Response：响应内容
```

6. 今日作业

```
【1】链家二手房爬虫
   # 注意：一切以响应内容为准
   1.1> 官网地址：进入链家官网，点击二手房 ： https://bj.lianjia.com/ershoufang/
   1.2> 目标 ： 抓取100页的二手房源信息，包含房源的：
        名称
        地址
        户型、面积、方位、是否精装、楼层、年代、类型
        总价
        单价
   1.3> 数据处理
        将数据分别存入：MySQL、MongoDB、csv文件中
```

```
12
13 【2】抓取快代理网站免费高匿代理，并测试是否可用来建立自己的代理IP池
14    https://www.kuaidaili.com/free/
```